Supplementary Methods

Microarray analysis

Analyses were performed in R version 2.9.0 (http://cran.rproject.org/). Annotated datasets are available at http://rock.icr.ac.uk/collaborations/Mackay/centroid.correlations.Eset. Annotated 'intrinsic' gene lists used for hierarchical clustering, clustered data, and Java Treeview files for each of the breast cancer datasets presented are available at http://rock.icr.ac.uk/collaborations/Mackay/observer.clustering.

Probe annotations and gene mapping of intrinsic gene lists

To overlay the intrinsic gene lists with the genes/probes of the microarrays from the different breast cancer datasets, the annotations of the intrinsic gene lists and breast cancer datasets were comprehensively updated and mapped to build 36 of the human genome (Ensembl assembly 54) [http://www.ensembl.org/index.html]) as described previously (1). Different types of gene identifiers (IDs), including Human Genome Organization (HUGO) gene symbols, Ensembl gene IDs and updated Unigene cluster IDs (http://www.ncbi.nlm.nih.gov/unigene), were used to annotate each intrinsic gene list and breast cancer dataset.

The intrinsic gene lists have generally been reported and annotated with Unigene cluster IDs. With each build of the Unigene database, however, UniGene clusters may be split into other clusters creating new Unigene IDs or may be retired. This results in probes in each microarray platform that carry multiple Unigene IDs, making unambiguous 1:1 matching of probes, genes, and Unigene IDs challenging. In order to address this problem, we updated each of the intrinsic gene lists and datasets used with updated Unigene annotation from build 218. Methodologies that rely on the overlay of features between different microarray platforms and datasets are more accurate and consistent if an ID more stable than Unigene cluster is chosen for the overlay.

The genes of the intrinsic gene lists published by Perou et al., 2000 (2), Sorlie et al., 2001 (3), Sorlie et al., 2003 (4), Hu et al., 2006 (5) and Parker et al., 2009 (6) were retrieved from the Stanford Microarray Database (http://smd.stanford.edu/) or

University of North Carolina Microarray Database (https://genome.unc.edu/cgi-bin/SMD/umad.pl).

The 496 probe intrinsic gene list (349 unique HUGO gene symbols) by Perou et al., 2000 (2) was annotated using SOURCE (http://smd.stanford.edu/cgi-bin/source/sourceSearch) based upon the HUGO gene symbols provided on the accompanying website of the original publication (http://genome-www.stanford.edu/breast_cancer/molecularportraits/index.shtml). The IMAGE clone IDs (http://image.hudsonalpha.org/) of the 456 probe intrinsic gene list (395 unique HUGO gene symbols) by Sorlie et al., 2001 (3), and the IMAGE clone IDs of the 552 probe intrinsic gene list (492 unique HUGO gene symbols) by Sorlie et al., 2003 (4), were updated and annotated for HUGO gene symbols using SOURCE and merged with Ensembl gene IDs using Entrez gene numbers. HUGO gene symbols, Unigene, and Ensembl IDs for the 1400 probe intrinsic gene list (1176 unique HUGO gene symbols) by Hu et al., 2006 (5) were retrieved directly from Biomart (www.ensembl.org/biomart/index.html) using the Agilent probe IDs supplied by the authors. The 1906 intrinsic gene list (1918 unique HUGO gene symbols) described by Parker et al., 2009 (6) was annotated using SOURCE based upon the HUGO gene symbols described in the supplementary information supplied by the authors. The probes/unique genes provided and identified for each intrinsic gene list and breast cancer dataset using HUGO gene symbols, Ensembl gene IDs and Unigene cluster IDs can be found in Supplementary Tables 2-7.

Probe annotations and gene mapping of breast cancer datasets

The publicly available normalized gene expression data of the NKI-295 (7) (n=295; http://microarray-pubs.stanford.edu/wound_NKI/explore.html), Wang (8) (n=286; GEO, accession number GSE2034) and TransBig (9) (n=198; GEO accession number GSE7390) datasets were retrieved. Gene annotations of these breast cancer datasets were updated based upon the individual probe identifier on each array used in the different studies to the current build of the genome NCBI36 Ensembl assembly 54 and Unigene build 218 using SOURCE, BioMart, and Matchminer (http://discover.nci.nih.gov/matchminer/index.jsp) as described previously (1).

Annotation of the U133A Affymetrix probe IDs to Ensembl gene IDs for the Wang

(8) and TransBig (9) datasets were retrieved directly from Ensembl assembly 54 using Biomart and merged with the publicly available array data supplied by the authors. The data for the NKI-295 breast cancers (7) were retrieved from the supplementary data published by Chang et al. (10). The 'Phil Green contig accession numbers' and 'Systemic Name' supplied by the authors were used as source of accession numbers to retrieve the HUGO gene symbol, Unigene cluster and Ensembl IDs using SOURCE, Biomart, and R as described above (1).

Datasets were then filtered for identifiers present in each intrinsic gene list using HUGO gene symbol, Unigene cluster, and Ensembl gene IDs. Of the identifiers tested, HUGO gene symbol was the annotation that allowed for the retrieval of the highest proportion of genes in the majority of intrinsic gene lists and datasets (Supplementary Table 7).

Hierarchical cluster analysis

The assignments of molecular subtypes of breast cancer based on hierarchical cluster analysis were essentially performed as previously described (2-6,11). As observed in the original dendrograms and descriptions of the intrinsic gene lists (2-5), when multiple probes mapped to the same gene, all were included in the hierarchical clustering analysis. Two-way average-linkage hierarchical clustering (median centered by feature/gene Pearson correlation as the gene similarity metric) was applied to each dataset using the software Cluster 3.0 (5,6,11). Cluster results were visualized using Java Treeview.

Heatmap figures of hierarchical cluster analyses

Dendrograms were generated as described above. A final heatmap for each intrinsic gene list for each dataset was produced. To recapitulate the heatmaps provided in the original publications, the defining genes/gene clusters as described in the text and/ or figures of the original publications were highlighted in the final dendrograms and heatmaps sent out to the observers and are summarized below:

1) **Perou et al., 2000 (2)**: Basal-like cluster including keratin 5 (KRT5) and keratin 17 (KRT17); Luminal/estrogen receptor (ER)+ cluster including estrogen receptor

1 (ESR1), GATA binding protein 3 (GATA3) and X-box binding protein 1 (XBP1); Human epidermal growth factor receptor 2 (HER2) cluster including v-erb-b2 erythroblastic leukemia viral oncogene homolog 2 (ERBB2); Normal breast-like cluster: given that no separate normal breast-like cluster but a second basal epithelial-cell-enriched gene cluster is shown in the original publication, and all normal breast-like samples consistently expressed pleiotrophin (PTN) as compared to basal-like, HER2, and luminal tumors (see Figure 3 in Perou et al. (2)), a PTN gene cluster was added.

2) **Sorlie et al., 2001 (3)**: Basal-like cluster including KRT5, KRT17 and fatty acid binding protein 7 (FABP7); Luminal cluster including ESR1, GATA3, trefoil factor 3 (TFF3) and XBP1; Luminal C/novel unknown cluster including squalene epoxidase (SQLE); HER2 cluster including ERBB2 and growth factor receptor-bound protein 7 (GRB7); Normal breast-like cluster including aquaporin 7 (AQP7), integrin, alpha 7 (ITGA7) and thrombospondin receptor (CD36).
3) **Sorlie et al., 2003 (4)**: Basal-like cluster including KRT5, KRT17 and cadherin 3, type 1, P-cadherin (CDH3); Luminal A cluster including ESR1, GATA3 and XBP1; Luminal B cluster including SQLE and lysosomal protein transmembrane 4 beta (LAPTM4B); HER2 cluster including ERBB2 and GRB7; Normal breast-like cluster including aldo-keto reductase family 1, member C1 (AKR1C1) and phosphoinositide-3-kinase, regulatory subunit 1 (PIK3R1).
4) **Hu et al., 2006 (5)**: Basal-like cluster including CDH3, forkhead box C1 (FOXC1), matrix metallopeptidase 7 (MMP7); Luminal cluster including ESR1, XBP1, GATA3 and TFF3; HER2 cluster including ERBB2 and GRB7; Interferon (IFN)-regulated cluster including signal transducer and activator of transcription 1 (STAT1); Proliferation cluster including topoisomerase (DNA) II alpha (TOP2A) and budding uninhibited by benzimidazoles 1 homolog (BUB1). It should be noted that no normal breast-like cluster is shown in the hierarchical cluster figure (Figure 2) and no genes characteristic of this molecular subtype are described in the text of the original publication. In addition, none of the characteristic genes described in Sorlie et al. 2001 (3), and Sorlie et al., 2003 (4), e.g. AQP7, ITGA7, CD36, AKR1C1, or PIK3R1 are part of this intrinsic gene set.
5) **Parker et al., 2009 (6)**: given that no specific gene clusters are shown in the hierarchical cluster figure (Fig A1) nor are genes characteristic for the five molecular subtypes described in the original publication, we inferred genes/gene

clusters in this intrinsic gene list from previous intrinsic gene list publications. Basal-like cluster including KRT17, CDH3 and MMP7; Luminal cluster including ESR1, GATA3 and XBP1; Luminal B cluster including SQLE and LAPTM4B; HER2 cluster including ERBB2 and GRB7; Normal breast-like cluster including AQP7, ITGA7 and CD36; Proliferation cluster including TOP2A and BUB1.

Molecular subtype assignment

To determine the reproducibility of microarray-based classification of breast cancers by hierarchical clustering analysis, the study curator (JSR-F) selected five researchers i) with experience in microarray-based expression profiling analysis, ii) previous publications on the use of the microarray-based molecular taxonomy of breast cancer, and iii) a first or senior author publication on microarrays in a journal with 2008 Thompson ISI impact factor greater than 5.

Guidelines that described how each molecular subtype should be identified by the visual analysis of the dendrograms obtained with hierarchical cluster analysis for each intrinsic gene list were sent to five of the authors (AM, BW, AG, BK, RN) via email, including the following:

1) A copy of the original studies describing the intrinsic gene lists and molecular subtype assignments by hierarchical clustering.
2) A separate copy of each of the hierarchical cluster analysis figures of the intrinsic gene lists of the original studies [i.e., Figure 3 from Perou et al. (2); Figure 1 from Sorlie et al. (3); Figure 1 from Sorlie et al. (4); Figure 2 from Hu et al. (5); Appendix Figure A1 from Parker et al. (6)].
3) 15 heatmap figures created using each of the five intrinsic gene lists of each of the three breast cancer datasets as shown in Supplementary Figures 1-15 (description see above).
4) An extract of the description of each of the molecular subtypes copied from the original publications:
    a) **General remarks**:
    - Perou et al., 2000 (2): “The two dendrogram branches in Fig. 3 largely separate the tumour samples into those that were clinically described as ER positive (blue) and those that were ER negative (other colours)”

- Sorlie et al., 2001 (3): ”The tumors were separated into two main branches. The left branch contained three subgroups previously defined (14). These groups all were characterized by low to absent gene expression of the ER and several additional transcriptional factors expressed in the luminal/ER+ cluster.”
- Sorlie et al., 2003 (4): ”The major distinction seen was between the tumors showing high expression of luminal epithelial specific genes including the ESR1 (Fig. 1G) and all other tumors showing low or no expression of these genes. The basal subtype (red) was the most homogeneous cluster of tumors, as reflected by the relatively short branches linking the tumors in this cluster (node correlation >0.4) and the deep branch separating it from the other subtypes (Fig. 1C).”; ”As in the Norway/Stanford data, the clearest discrimination was between tumors that expressed genes in the luminal A/ESR1 cluster at high levels (Fig. 2C) and the tumors that were negative for these genes and exhibited expression profiles characteristic of either the basal, the ERBB2+ or the luminal B subtypes (Fig. 2 D–F).”
- Parker et al., 2009 (6): ”Significant clusters representing the "intrinsic" subtypes luminal A (LumA), luminal B (LumB), HER2-enriched, basal-like, and normal-like.”

b) **Basal-like**:

- Perou et al., 2000 (2): “All six of these tumours showed staining for either keratins 5/6 or 17 or both (Fig. 2d). Notably, these six tumours also failed to express ER and most of the other genes that were usually co-expressed with it (Fig. 3c).”
- Sorlie et al., 2001 (3): “The basal-like subtype (Fig. 1 A, red) was characterized by high expression of keratins 5 and 17, laminin, and fatty acid binding protein 7 (Fig. 1 E).”
- Hu et al., 2006 (5): “A Basal-like expression cluster was also present and contained genes (i.e. c-KIT, FOXC1 and P-Cadherin) previously identified to be characteristic of basal epithelial cells (Figure 2F).”

c) **HER2**:

- Perou et al., 2000 (2): “Overexpression of the Erb-B2 oncogene was associated with the high expression of a specific subset of genes. We

identified a cluster of tumours that was partially characterized by the high level of expression of this subset of genes (Fig. 3d). These tumours also showed low levels of expression of ER and of almost all of the other genes associated with ER expression—a trait they share with the basal-like tumours."

- Sorlie et al., 2001 (3): "The ERBB2+ subtype (Fig. 1 A, pink) was characterized by high expression of several genes in the ERBB2 amplicon at 17q22.24 including ERBB2 and GRB7 (Fig. 1 C)."
- Hu et al., 2006 (5): "As shown in previous studies, a HER2+ expression cluster was observed in the cluster analysis of the "combined test set" and contained multiple genes from the 17q11 amplicon including HER2/ERBB2 and GRB7 (Figure 2D). The HER2+ intrinsic subtype (pink dendrogram branch in Figure 2B) was predominantly ER-negative (i.e. HER2+/ER-) as previously shown."

d) **Luminal**:

- Perou et al., 2000 (2): "The tumours in the ER+ group were characterized by the relatively high expression of many genes expressed by breast luminal cells (Fig. 3c). This connection was further corroborated using immunohistochemical analysis and antibodies against the luminal cell keratins 8/18 (Fig. 2c). With one exception, none of the tumours in this group expressed Erb-B2 at high levels (Fig. 3d)."
- Hu et al., 2006 (5): "As shown in previous studies, a Luminal/ER+ expression cluster was present and contained ER, XBP1, FOXA1 and GATA3 (Figure 2C). GATA3 has recently been shown to be somatically mutated in some ER+ breast tumors, and some of the genes in Figure 2C are GATA3-regulated (FOXA1 and TFF3), thus showing the functional clustering of a transcription factor and some of its direct targets."

e) **Luminal A**:

- Sorlie et al., 2001 (3): "The group of 32 tumors (termed luminal subtype A, Fig. 1 A, dark blue) demonstrated the highest expression of the ER α gene, GATA binding protein 3, X-box binding protein 1, trefoil factor 3, hepatocyte nuclear factor 3 α, and estrogen-regulated LIV-1 (Fig. 1 G)."

f) **Luminal B**:

- Sorlie et al., 2001 (3): “The second group of tumors positive for luminal-enriched genes could be broken into two smaller units, a small group of five tumors termed luminal subtype B (Fig. 1 A, yellow) and the group of 10 tumors called luminal subtype C (Fig. 1 A, light blue). Both of these groups showed low to moderate expression of the luminal-specific genes including the ER cluster.”

g) **Luminal C**:

- Sorlie et al., 2001 (3): “The second group of tumors positive for luminal-enriched genes could be broken into two smaller units, a small group of five tumors termed luminal subtype B (Fig. 1 A, yellow) and the group of 10 tumors called luminal subtype C (Fig. 1 A, light blue). Both of these groups showed low to moderate expression of the luminal-specific genes including the ER cluster.”; “Luminal subtype C was further distinguished from luminal subtypes A and B by the high expression of a novel set of genes whose coordinated function is unknown (Fig. 1 D), which is a feature they share with the basal-like and ERBB2+ subtypes.”

h) **Normal breast-like**:

- Perou et al., 2000 (2): “Several tumour samples and the single fibroadenoma tested (Fig. 3, light green), were clustered with a group of samples that also contained the three normal breast specimens (Fig. 3a). The 'normal breast' gene expression pattern is typified by the high expression of genes characteristic of basal epithelial cells and adipose cells, and the low expression of genes characteristic of luminal epithelial cells.”
- Sorlie et al, 2001 (3): “Tumor samples included in the normal breast-like group (Fig. 1 A, green) showed the highest expression of many genes known to be expressed by adipose tissue and other nonepithelial cell types (Fig. 1 F). These tumors also showed strong expression of basal epithelial genes and low expression of luminal epithelial genes.”

i) **Interferon (IFN)-regulated cluster**:

- Hu et al., 2006 (5): “A possible new tumor group (IFN) characterized by the high expression of Interferon (IFN)-regulated genes was observed in the combined test set analysis (Figure 2E). According to EASE, the GO categories "immune response" and "defense response" were over-represented relative to

chance in the interferon-regulated gene cluster. This cluster contained STAT1, which is thought to be the transcription factor responsible for mediating IFN-regulation of gene expression."

j) **Proliferation cluster**:

- Hu et al., 2006 (5): "The most significant difference between the previous Intrinsic/Stanford gene lists and the new Intrinsic/UNC gene list was that the latter contained a large proliferation signature (Figure 2G)."

Observers were requested to classify each dataset according to the methods described by Perou et al. (2), Sorlie et al. (3), Sorlie et al. (4), Hu et al. (5), and Parker et al. (6), identifying all molecular subtypes described in each publication. If samples could not be assigned to a molecular subtype with confidence, the observers could opt for considering the sample as unclassifiable, as previously done in Sorlie et al. (4) and Parker et al. (6). A request to keep the correspondence strictly confidential was made. No discussions with other researchers were permitted. The identity of each observer was kept confidential to the other study participants. Molecular subtype assignments were made by each observer blinded to the results reported by the other observers and sent directly to the study curator.

Supplementary References

1. Weigelt B, Mackay A, A'hern R, et al. Breast cancer molecular profiling with single sample predictors: a retrospective analysis. *Lancet Oncol*. 2010;11(4):339-349.
2. Perou CM, Sorlie T, Eisen MB, et al. Molecular portraits of human breast tumours. *Nature*. 2000;406(6797):747-752.
3. Sorlie T, Perou CM, Tibshirani R, et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci U S A*. 2001;98(19):10869-10874.
4. Sorlie T, Tibshirani R, Parker J, et al. Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc Natl Acad Sci U S A*. 2003;100(14):8418-8423.
5. Hu Z, Fan C, Oh DS, et al. The molecular portraits of breast tumors are conserved across microarray platforms. *BMC Genomics*. 2006;7:96.

6. Parker JS, Mullins M, Cheang MC, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. *J Clin Oncol*. 2009;27(8):1160-1167.
7. van de Vijver MJ, He YD, van't Veer LJ, et al. A gene-expression signature as a predictor of survival in breast cancer. *N Engl J Med*. 2002;347(25):1999-2009.
8. Wang Y, Klijn JG, Zhang Y, et al. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet*. 2005;365(9460):671-679.
9. Desmedt C, Piette F, Loi S, et al. Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the TRANSBIG multicenter independent validation series. *Clin Cancer Res*. 2007;13(11):3207-3214.
10. Chang HY, Nuyten DS, Sneddon JB, et al. Robustness, scalability, and integration of a wound-response gene expression signature in predicting breast cancer survival. *Proc Natl Acad Sci U S A*. 2005;102(10):3738-3743.
11. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A*. 1998;95(25):14863-14868.